**Abstract**

Rising global temperatures may in turn lead to more intense wildfires. Smoke from wildfires can also pose serious health risks and damage homes and businesses.

This project looks at how data science can help predict how intense wildfires will be, how long they might last, and where they could go. We are using data from the USDA, NASA, and weather services to build a model that can help emergency responders prepare for and manage wildfires.

**Outline of the Project**

* Business Purpose: Build a model that predicts wildfire danger to help emergency teams and government agencies respond better.
* Practical Application: Help with wildfire preparedness and response using data predictions.
* Importance: Wildfires are harmful to the environment, health, and economy. A reliable model could help reduce these impacts.
* Stakeholders: Firefighters, scientists, government officials, mentors, and sponsors.
* Assumptions:
  + Fire intensity and emissions can be predicted using weather and environmental data.
  + Data can be matched by location and time.
* Scope:
  + *Included:* Predicting fire strength, duration and direction. Combining multiple datasets.
  + *Excluded:* Real-time fire alerts or new hardware systems.

**Success Criteria**

* SC1: Successfully complete data merging, cleaning, and feature engineering.
* SC2: Build and test at least one machine learning model that can provide basic predictions or insights about wildfire behavior.

**Data Assumptions and Limitations**

* We first thought wind direction data was missing, but we found it.
* Weather and emissions data are not always recorded at the same time or place, so we use a nearest-neighbor method to link them.
* We are currently having trouble merging the weather and emissions data due to limited RAM in our Google Cloud environment. This issue has delayed our progress. Completing data merging, cleaning, and feature engineering would help us move forward more efficiently if delays continue.

**Summary of Data Processing and Aggregation**

* Target Variables: CO2 Emissions and Area Burned from USDA
* Predictors: Land type, fuel type, wind speed/direction, heat index, humidity, temperature, etc.
* Aggregation: We used a spatial join to match data by location and grouped it by day.
* Preprocessing: We scaled the numbers, grouped some outputs into categories, and converted text fields into numbers.

**Data Visualizations (Using CSV files in google drive)**

1. Burn Index Histogram: Shows most areas are low-risk, with a second group at medium risk.

A graph of a distribution of burn index

AI-generated content may be incorrect.

1. Area Burned Histogram: Most records show small fires; a few are very large.

A graph of distribution of area burned

AI-generated content may be incorrect.

1. Wildfire Locations Map: High fire activity in the southeast U.S.

A map of the united states

AI-generated content may be incorrect.

1. Boxplot of CO Concentration: Some very high values make the data uneven.

A graph with numbers and lines

AI-generated content may be incorrect.

1. Coverage Map: Confirms our data covers most of the U.S.

A blue map of the united states

AI-generated content may be incorrect.

1. Spatial Join: Shows how we linked weather and emissions data using location.

A map of the united states

AI-generated content may be incorrect.

**Summary of Modeling and Analysis**

We are currently focused on data processing and have not yet started building models. The main tasks in progress include merging emissions and weather data, cleaning and transforming the variables, and performing exploratory analysis. Modeling and evaluation will begin once we complete the data preparation phase. The following are being considered:

* Feature Importance Analysis: To identify which weather or emission features most influence fire behavior.
* Clustering (e.g., KMeans): To explore natural groupings or patterns in fire activity without labels.
* Precision & Recall: Useful when wildfire events are rare and the dataset is imbalanced. Precision helps reduce false alarms, while recall helps avoid missing real fire events.
* F1 Score: Balances precision and recall, especially useful when dealing with imbalanced classes.
* Accuracy: Not a reliable measure for imbalanced data. High accuracy can be misleading if the model simply predicts the dominant class.
* AUC-ROC: Helps evaluate how well the model ranks predictions in binary classification tasks. Less useful for multi-class scenarios.
* Confusion Matrix: Provides a complete view of prediction performance across different fire intensity levels, highlighting where the model succeeds or fails.

**Future Work Plan**

|  |  |
| --- | --- |
| Task Description | Date |
| Finish cleaning and adjusting input features | Week 2 |
| Choose best model and evaluation tools | Week 3 |
| Train models with updated data | Week 4 |
| Tune model settings for best performance | Week 5 |
| Look at how fire risk changes over time | Week 6 |
| Build charts and finish final report | Week 7–8 |

**Potential Concerns [C] and Blockers [B]**

* B1: Limited RAM in the current cloud setup is delaying data merging. We may need additional computing resources to continue efficiently.
* B2: We might need more data processing power for better predictions.
* C1: We have more data for small fires than large ones.
* C2: Time and location differences between datasets can be hard to align.
* C3: Running complex models could require more computing power.